iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** X **Month of publication:** October 2018

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue X, Oct 2018- Available at www.ijraset.com

Application of K-means Clustering Technique for Analysis of Students Academic Performance in School Education

Antariksh Pandey[1], Neelam Sahu[2], Ayush Kumar Agrawal[3]
[1, 2, 3] *Department of IT, Dr C V Raman Univirsity*

Abstract: ***The critical issue to the academic community is the ability to monitor the progress of students' academic performance. The status or image of any institute is specifically depends upon of students' academic performance. This paper introduced a model to categorized students of a class on the basis of their academic performances. A data of 40 student of class 11th is used for the experimental work. K-means clustering algorithm is used for analyzing students' result data. The results are obtains in the form of clusters through which institution can easily trace their average or week students and make some more efforts on their studies.***

Keywords: Academic performance, Average percentage, Overall performance score, K-means, Cluster.

I. INTRODUCTION

Average percentage making of all subjects is commonly used as an indicator of academic performance of any student in many schools. Many universities set a minimum average percentage that should be maintained in order to take admission in the degree program. In many universities, the minimum percentage set for the student is 60 to take admission for any graduate program. Average percentage of 70 and above is considered an indicator of excellent academic performance. Therefore, Average percentage still remains the most common factor used by the academic planners to evaluate progression in an academic environment [1]. Many factors could act as barriers to students attaining and maintaining a high, Average percentage that reflects their overall academic performance during their tenure in academic institutions. These factors could be targeted by the faculty members in developing strategies to improve student learning and improve their academic performance by way of monitoring the progression of their performance. Therefore, performance evaluation is one of the bases to monitor the progression of student academic performance in school. Base on this critical issue, grouping of students into different categories according to their subject making or average percentage has become a complicated task.

Some of data mining techniques, such as clustering algorithm, the discovery of key characteristics from the students' performance is possible and possibly use those characteristics for future prediction. Some promising results from applying k-means clustering algorithm have been achieved. There have been some promising results from applying k-means clustering algorithm with the Euclidean distance measure, where the distance is computed by finding the square of the distance between each scores, summing the squares and finding the square root of the sum [2].

Cluster analysis could be divided into hierarchical clustering and non-hierarchical clustering techniques. Examples of hierarchical techniques are single linkage, complete linkage, average linkage, median, and Ward. Non-hierarchical techniques include k-means, adaptive k-means and fuzzy clustering. To determine which algorithm is good is a function of the type of data available and the particular purpose of analysis. In more objective way, the stability of clusters can be investigated in simulation studies [3]. The problem of selecting the "best" algorithm/parameter setting is a difficult one. A good clustering algorithm ideally should produce groups with distinct non-overlapping boundaries, although a perfect separation can't typically be achieved in practice. Figure of merit measures (indices) such as the silhouette width or the homogeneity index [4] can be used to evaluate the quality of separation obtained using a Clustering algorithm. The concept of stability of a clustering algorithm was considered in [5]. The idea behind this validation approach is that an algorithm should be rewarded for consistency. In this paper, we implemented traditional K-means clustering algorithm [6] was chosen to be used in the analysis of the students' scores.

In this paper we use k-means clustering algorithm as a simple and efficient tool for the analysis of students' academic performance is school. The aim of this research is to produce a model for academic institutions to find their best, average and poor student with respect to their academic performances. This system helps in decision making for academic institutions to make some or more efforts towards students learning.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue X, Oct 2018- Available at www.ijraset.com

II. METHODOLOGY

K-Means clustering algorithm developed [7] three decades ago is one of the K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the most popular clustering algorithm used in variety of domains. A prior knowledge of number of clusters are must for K-means clustering algorithm. K-Means is defined over continuous data[8]. K-Means algorithm calculates its centers iteratively[9]. Let D ={d_i| 1,....n} be a data set having K-clusters, C={c_i| 1,....k} be a set of K centers and Sj= {d | d is the o of cluster k} is member of cluster kg be the set of samples that belong to the k^{th} cluster. K-Means algorithm minimizes the following function which is defined as a cost function

$$\text{Cost} = \sum_{i=1}^{n} \text{dist}(di, ck) \quad (1)$$

where dist(di,ck) measures the Euclidean distance between a pattern di and its cluster center ck .
nearest mean. K-means starts with a single cluster with its center as the mean of the data. This cluster is split to two and the means of the new clusters are trained iteratively. These clusters again split and the process continues until the specified number of clusters is obtained. If the specified number of clusters is not a power of two, then the nearest power of two above the number specified is chosen. Then the least important clusters are removed and the remaining clusters are again iteratively trained to get the final clusters [10].
Generalised Pseudocode of Traditional k-means is as follows [11][12]:

1) Step 1: Accept the number of clusters to group data into and the dataset to cluster as input values
2) Step 2: Initialize the first K clusters
a) Take first k instances or
b) Take Random sampling of k elements
3) Step 3: Calculate the arithmetic means of each cluster formed in the dataset.
4) Step 4: K-means assigns each record in the dataset to only one of the initial clusters
a) Each record is assigned to the nearest cluster using a measure of distance (e.g Euclidean distance).
5) Step 5: K-means re-assigns each record in the dataset to the most similar cluster and re-calculates the arithmetic mean of all the clusters in the dataset.

In this research we use the application of K-means technique to find cluster and then calculate average percentage of each cluster for the analysis of overall performance of students in a cluster. The performance index which is presented in table I is used to predict the performance students in cluster. Better results are obtained using changes in no of clusters in k means method.

TABLE I
PERFORMANCE INDEX

Average Percentage range	Performance of Student
Above 70	Excellent
60-70	Very Good
50-60	Good
40-50	Fair
Bellow 40	Poor

III.EXPERIMENTAL WORK AND RESULT ANALYSIS

For analysing the academic performance of students, data of 40 student of class 11th (Maths stream) of Vedic Convent School, Bilaspur (CG), is taken presented in Table II. The data contain 7 attributes where first column have name of all 40 student, next five columns contain makes of students in Maths, Physics, Chemistry, Hindi and English subject respectively and the last 7th column contain the total average percentage of each student. K-means clustering technique is then implemented in this data for the grouping of different student with respect to their performances in last academic year.

TABLE III
11^{TH} CLASS DATA OF 40 STUDENT

S. No	Student Name	Marks Obtained					Average Percentage
		Maths	Physics	Chemistry	Hindi	English	
1	S1	44	56	52	58	60	54
2	S2	80	88	90	84	92	86.8
3	S3	60	79	79	79	77	74.8
4	S4	59	66	63	68	64	64
5	S5	40	50	49	50	51	48
6	S6	70	69	66	63	60	65.6
7	S7	69	77	73	77	78	74.8
8	S8	62	77	80	79	77	75
9	S9	64	77	78	80	76	75
10	S10	86	88	84	83	80	84.2
11	S11	50	57	60	64	63	58.8
12	S12	60	75	70	77	88	74
13	S13	60	74	69	70	75	69.6
14	S14	57	69	67	69	70	66.4
15	S15	69	77	77	76	79	75.6
16	S16	57	59	60	65	67	61.6
17	S17	61	76	78	80	76	74.2
18	S18	51	59	54	64	66	58.8
19	S19	65	66	69	70	72	68.4
20	S20	67	70	61	64	65	65.4
21	S21	54	60	62	66	69	62.2
22	S22	57	70	72	74	68	68.2
23	S23	55	64	68	69	68	64.8
24	S24	69	89	88	84	76	81.2
25	S25	54	67	68	70	65	64.8
26	S26	65	77	79	83	86	78
27	S27	71	89	87	91	86	84.8
28	S28	60	71	77	74	79	72.2
29	S29	40	67	61	54	55	55.4
30	S30	53	60	65	61	62	60.2
31	S31	47	56	57	60	67	57.4
32	S32	61	67	69	75	74	69.2
33	S33	62	69	71	75	67	68.8
34	S34	78	79	84	82	88	82.2
35	S35	45	56	59	61	56	55.4
36	S36	60	76	78	75	77	73.2
37	S37	40	66	56	58	61	56.2
38	S38	61	66	69	70	70	67.2
39	S39	44	40	60	56	56	51.2
40	S40	70	68	71	78	77	72.8

The experimental work is done with two cases on the basis of no of cluster decided initially.

Case 1: In this case the no of cluster is set as 3 initially. The no of cluster with silhouette value after implementation of K means is shown in figure 1. Table III presents the cluster size and overall performance of students in any cluster. The size of cluster 1 ,2 and 3

is 16, 14 and 10 respectively. The overall performance of cluster 1, cluster 2 and cluster 3 is 77.42, 66.15 and 55.54. The graphs are generated in figures 3, where the overall performance is plotted against the cluster size. Through the results it is clearly seen that 16 out of 40 students are performing "Excellent" in their studies. As same 14 out of 40 students are "Very Good" in their studies. But 10 out of 40 students are just "Fair". But this cluster configuration is not appropriate because most objects have low value.

TABLE IIIII
OBTAINED RESULTS FOR CLUSTERS

Cluster#	Cluster size	Overall Performance
1	16	77.42
2	14	66.15
3	10	55.54

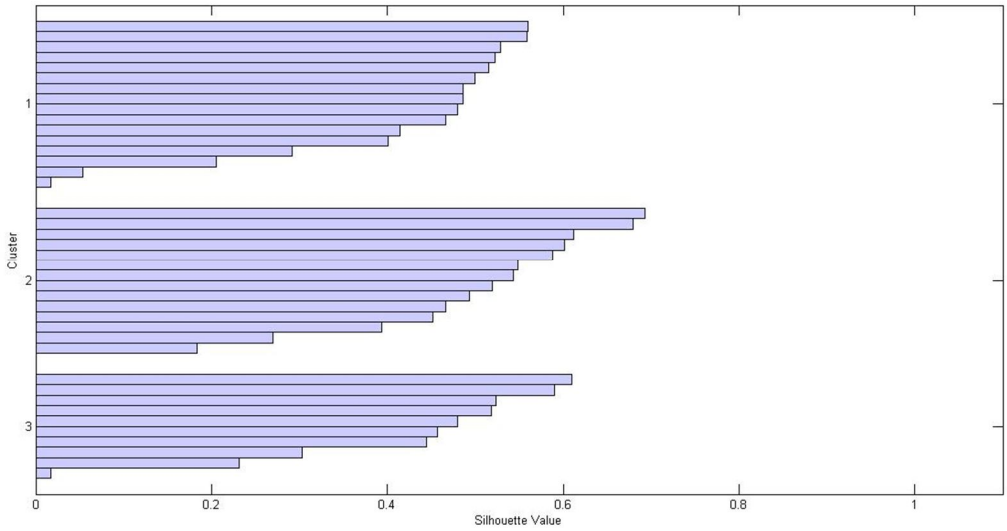


Fig. 1 Different silhouette value for each object of cluster

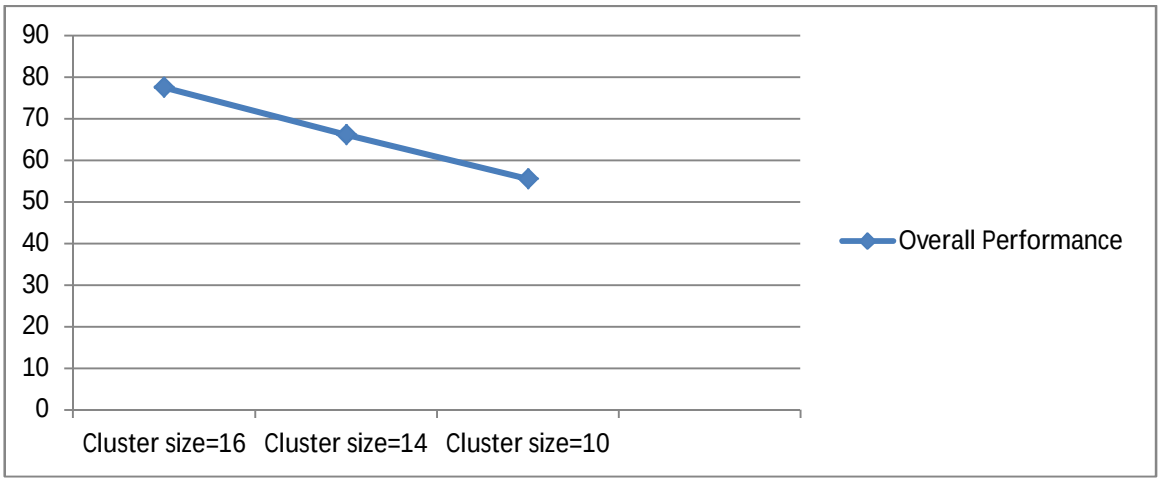


Fig. 2 Overall performance of each cluster

Case 2: In this case the no of cluster is set as 4 initially. The no of cluster with silhouette value after implementation of K means is shown in figure 3. Table IV presents the cluster size and overall performance of students in any cluster. The size of cluster 1, 2, 3 and 4 is 10, 5, 14 and 11 respectively. The overall performance of cluster 1, cluster 2, cluster 3 and cluster 4 are 55.54, 83.84, 66.15 and 74.5. This cluster configuration is highly appropriate because maximum objects have higher value with comparison to what we get in case 1. The graphs are generated in figures 4, where the overall performance is plotted against the cluster size. Through the

results it is clearly seen that 16 out of 40 students are performing "Excellent" in their studies. As same 14 out of 40 students are "Very Good" in their studies. But 10 out of 40 students are just "Fair".

TABLE IV
OBTAINED RESULTS FOR CLUSTERS

Cluster#	Cluster size	Overall Performance
1	10	55.54
2	5	83.84
3	14	66.15
4	11	74.5

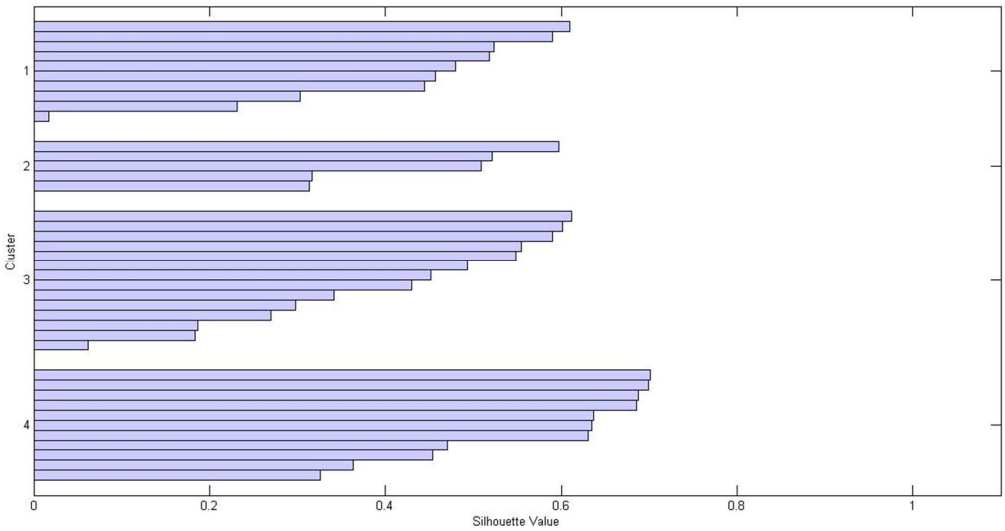


Fig. 3 Different silhouette value for each object of cluster

Fig. 4 Overall performance of each cluster

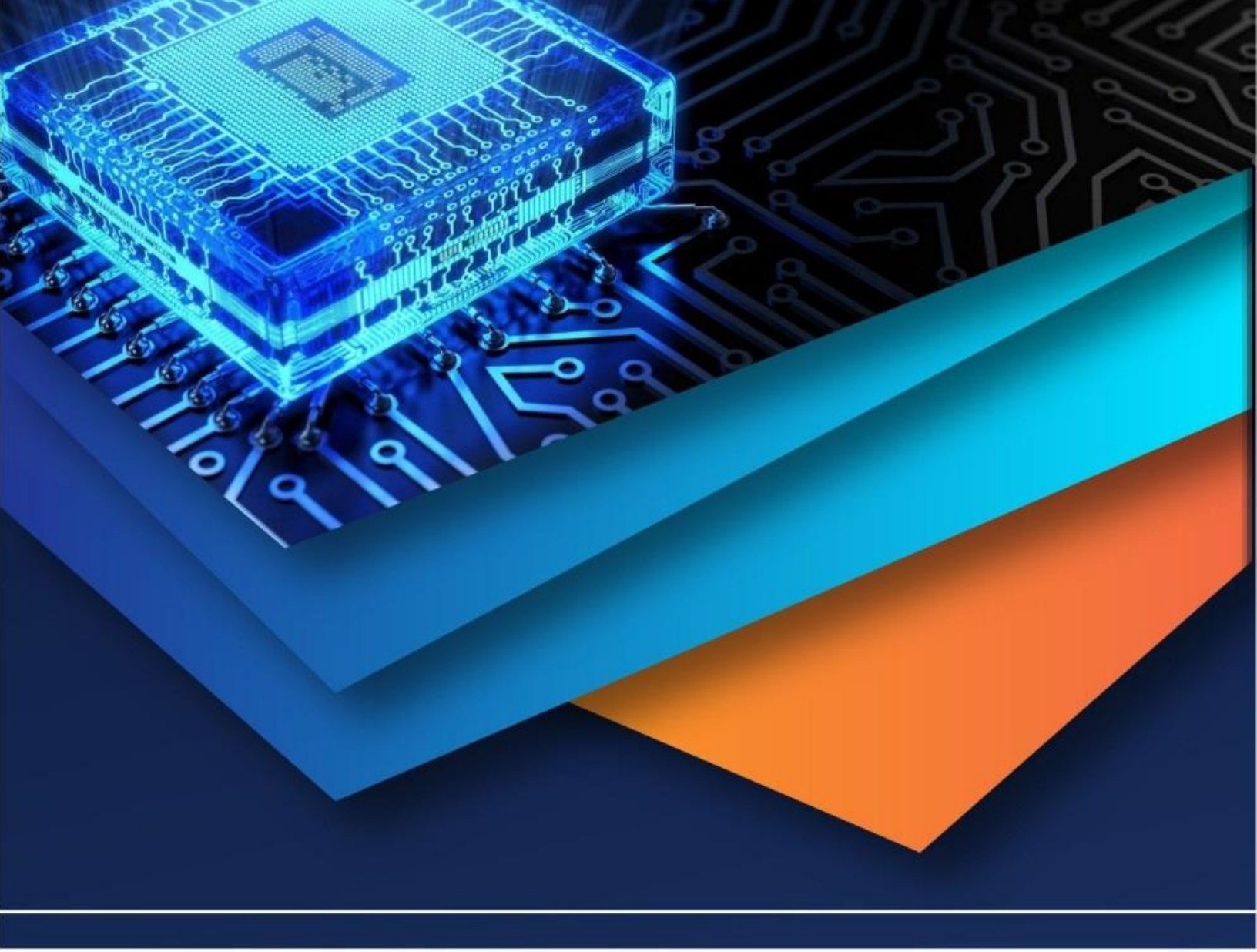
It is clearly seen from Table III that cluster 4 have 11 objects or student whose performances are "Fair" that indicates that those students are not up to the mark as other students. So through this result school can traces those students and can do some extra work in field of student learning. This effort may be in the type of extra classes by subject expert teachers or using so effective turning techniques.

IV. DISCUSSION AND CONCLUSION

Analysis of student's academic performance is the major issue for the status and reputation of any academic institution. This analysis helps in decision making for academic institutions to improve their education techniques. A common approach is to use average percentage to evaluate the academic performance of any student. In this research we introduced a model to evaluate students' academic performance. We use the application of K- means that creating no of clusters and the find the performance of students of every cluster using performance index. A cluster with 'Fair' or 'Poor' performances indicates the academic institution to make effective changes to their education techniques for those students. In this research K-means method is applied with initialization of no of cluster is 3 and 4. As results shown above the process is appropriate when cluster is 4. The performance of cluster 1 is 'Fair' that indicates that institution should take care of those students.

REFERENCES

[1] S. Sujit Sansgiry, M. Bhosle, and K. Sail, "Factors that affect academic performance among pharmacy students," American Journal of Pharmaceutical Education, 2006.
[2] A. M Fahim., A. M Salem., F. A Torkey and M. A. Ramadan, "An efficient enhanced k-means clustering algorithm," Journal of Zhejiang University Science A., pp. 1626–1633, 2006
[3] R. Sharmir and R. Sharan , "Algorithmic approaches to clustering gene expression data," In current Topics in Computational Molecular Biology MIT Press; pp. 53-65, 2002.
[4] H. J. Mucha , "Adaptive cluster analysis, classification and multivariate graphics,"Weirstrass Institute for Applied Analysis and Stochastics, 1992.
[5] P. J Rousseeuw, "A graphical aid to the interpretation and validation of cluster analysis," Journal of Computational Appl Math, vol 20, pp. 53– 65, 1987.
[6] A.M. Fahim , A.M. Salem , F.A. Torkey and M.A. Ramadan , "An efficient enhanced k-means clustering algorithm," Journal of Zhejiang University Science A., pp. 1626–1633, 2006.
[7] Mac Queen J., "Some methods for classification and analysis of multivariate observations" Proc. 5th Berkley Symp. On Mathematical Statistics and Probability, Vol 1, pp. 281–297,1967
[8] Fukunaga, K., " Introduction to Statistical Pattern Recognition. Academic Press, 1990.
[9] Gersho A., Gray R.M., "Vector Quantization and Signal Compression", KAP, 1992.
[10] Nanda S.R. , Mahanty B. and Tiwari M.K., "Clustering Indian stock market data for portfolio management", Expert Systems with Applications, .vol 37, pp. 8793–8798, 2010.
[11] Oyelade, O. J, Oladipupo, O. O and Obagbuwa, I. C, "Application of k-Means Clustering algorithm for prediction of Students' Academic Performance", International Journal of Computer Science and Information Security (IJCSIS), Vol. 7, 2010.
[12] Shehroz and S. K. and Ahmad A., "Cluster center initialization algorithm for K-means clustering", Pattern Recognition Letter, vol. 25, pp. 1293–1302, 2004.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY